QUANTIFICATION AND GENERALIZATION IN LINGUISTIC THEORY

S.Rajendran

rajushush@gmail.com

Quantitative linguistics studies the frequency and distribution of linguistic units using statistical techniques. The subject has both pure and applied side: the former aims to establish general principles concerning the statistical regularities governing ways words, sounds, etc. are used; the latter investigates the ways statistical techniques can be used to elucidate linguistic problems (such as functional load, stylistic distinctiveness, authorship identity). Considerable use is made of corpora. Productivity which is closely related to generativity refers the creative capacity of language users to produce and understand and identify a large number of sentences, phrases or words. A pattern is productive if it is repeatedly used in language to produce further instance of the same type. Non-productive patterns lack any such potential. Quantitative analysis sometimes tries to capture the productive capacity of a particular principle by quantification of the resultant occurrence of forms produced by the principle. It goes to the extent of capturing any linguistic generalization by means of statistical information. Many computational linguists make use of statistical techniques to capture linguistic generalizations. Quantification helps computational linguists in many ways, for examples it helps them in fixing ambiguous grammatical categories (categorical disambiguation), ambiguous patterns (pattern disambiguation), ambiguous words (word sense disambiguation) and so on. Field linguistics has taught linguists the reliability of corpus for language analysis. The present day corpus linguistics which rely on quantitative data, has evolved as an effective and dependable methodology to arrive at generalization to promote various applications. Though the controversy over empiricism and rationalism for linguistic analysis is a good old issue, it is valid even in the current status of linguistics as a hybridized discipline.

For the application of linguistics to process data amenable to computation, it hangs on the extreme formalism propounded by mathematical linguistics which could be claimed as nonquantitative discipline on one side and quantification promoted by quantitative linguistics on the other side. The present paper addresses the reliability of mathematical and statistical techniques in arriving at linguistic generalization.

Argument for and against quantification

Taking frequency element as a component of language is a controversial question. There are linguist who maintain that frequency of linguistic form is not a part of language but belongs only to the use of words which is quite arbitrary. The advocates of quantification argue that the traditional linguist overlook the difference between langue, the abstraction (language) and the parole, the actual

instances of the language. It is an empirical fact, steadfastly ignored by the opponents of the frequency concept in language, that the frequency of linguistic forms shows a stability decidedly comparable to that of the traditional linguistic analysis.

The opponents of quantitative analysis argue that statistical methods are not linguistically relevant. When statistical methods were first applied to language, the usual or typical objection was that language was all deterministic and had nothing to do with chances. Promoters of quantitative linguistics argue that it was proved that chances, as well as choice, had a big part to play in language. Chomsky raised the objection that so far as chance did play a role in language, this aspect was only what language had in common with many other part of the physical universe, and thus was 'linguistically not relevant'. Chomsky set aside any theory of grammar which does not take care of the generative aspect as of marginal importance. Quantitative linguist contends that it is precisely the statistical nature of certain linguistic variable that provides the basis for what they call stylostatistics, and this is eminently useful in language research. Chomsky contends that only a statistical model with 'very surprising' properties accounting for some statistical regularity is of interest to the linguist. The quantitative linguists criticizes Chomsky's objection to statistical generalization by claiming that Chomsky completely forgets the practical aspect of establishing statistical regularities. Whether they are 'very' or 'less surprising', they enable us to make predictions, and thus to know what to expect under normal circumstances. In statistical linguistics, it is this which forms the basis of stylo-statistical work, since it provides the statistical yardstick against which to measure observed differences between writers for significance.

Linguistic rules and statistical interpretation

Linguists prefer to describe the form of a language in terms of rules. They resort to explain the distribution of linguistic units in terms of rules. For them the linguistic units are amenable to 'item and arrangement' or 'item and process'. They posit rules to capture the generality in distribution. The items are arranged depicting underlying patterns or rules. They rely on immediate constituents, collocations, selectional restrictions and neatness of pattern in depicting the structure of a language.

The proponents of statistical linguistics argue that the Saussurean dichotomy of significant and signifie, lends ways to make the structure amenable to numerical methods. Grammar, in its traditional form, may be regarded as the prototype of linguistic structure. It follows that, to grammar, structuralism is at least supererogatory, since grammar is structure. Grammar, although being system of linguistic generalities as against lexical peculiarities, cannot be regarded as a system implying complete independence of linguistic form and meaning. The development of structuralism would seem to be towards a conception of language affording higher and higher degree of independence of

form and meaning. This makes the vocabulary amenable to statistical parameters.

One may ask the following question: What has been achieved by the introduction of frequency element for the purpose of offering a higher degree of structuralism? The answer given by the quantitative linguists goes in the following fashion: On the phonemic level, structuralism was a matter of independence of sound and meaning of individual words; on the vocabulary level, it is a matter of the independence of the frequency distribution of linguistic forms for the content of literary texts. Since continuous texts differ in the grammatical arrangement of vocabulary items as well as, to some extent, in these items themselves, the statistical concept of the use of words can be regarded as higher degree of structuralism compared with grammar.

Word order, as the sequence of linguistic symbols within the sentence, is clearly a feature which does not reside in the manifold of experience; it does not lie in the physical or psychological objects but arises only through expression in language. This suggests a statistical law as governing word order in question.

Empiricism vs. rationalism

In the early days, the linguist depends on the data which is a small form of corpus in arriving at the structure of a language. This approach, as it is based on observation and deduction, is branded as empirical approach. Chomsky in a series of influential publications, changed the direction of linguistics away form empiricism and towards rationalism in a remarkably short period of time. A rationalist theory is based on artificial behavioural data, and conscious introspective judgements. This may be a native speaker of a language reflecting on that language and making theoretical claims based on those reflections. Rationalist theories are based on the development of a theory of mind in the case of linguistics, and have as fundamental goal cognitive plausibility. The aim is to develop a theory of language that not only emulates the external effects of human language process, but actively seeks to make the claim that it represents how the processing is actually undertaken. On the other hand, an empiricist approach to language is dominated by the observation of naturally occurring data, typically through the medium of the corpus.

Chomsky argues that the goals of linguistics are not the enumeration and description of performance phenomena, but rather they are introspection and explanation of linguistic competence. If we accept the enumeration and description as a goal for linguistics, it seems an unattainable one, as natural languages are not finite. As a consequence, the enumeration of sentences can never possibly yield an adequate description of language. How can a partial corpus be the sole explicadum of an infinite language? We must not eschew introspection entirely. If we do so, detecting ungrammatical structures and ambiguous structures becomes difficult, and indeed may be impossible. The

power and compulsion of these arguments ensured the almost total rejection of corpus-based (empirical) methodologies in linguistics and the establishment of new orthodoxy. Introspection (rationalist) based approaches rose in the ascendant.

The importance of copora to language study is aligned to the importance of empirical data. Empirical data enable the linguist to make objective statements, rather than those which are subjective, or based upon the individual's own internalised cognitive perception of language. Empirical data also allows us to study language varieties such as dialects or earlier periods in a language for which it is not possible to carry out a rationalist approach. It is important to note that although many linguists may use the term "corpus" to refer to any collection of texts, when it is used here it refers to a body of text which is carefully sampled to be maximally representative of the language or language variety. Corpus linguistics, proper, should be seen as a subset of the activity within an empirical approach to linguistics. Although corpus linguistics entails an empirical approach, empirical linguistics does not always entail the use of a corpus.

Quantitative versus non-quantitative approaches

Statistics is known to be a quantitative approach to research. However, most of the research done in the fields of language and linguistics is of a different kind, namely qualitative. Succinctly, qualitative analysis differs from quantitative analysis is that in the former no attempt is made to assign frequencies, percentages and the like, to the linguistic features found or identified in the data. In quantitative research, linguistic features are classified and counted, and even more complex statistical models are constructed in order to explain these observed facts. In qualitative research, however, we use the data only for identifying and describing features of language usage and for providing real occurrences/examples of particular phenomena. We can show how quantitative methods and statistical techniques can supplement qualitative analyses of language. We can present some mathematical and statistical properties of natural languages, and introduce some of the quantitative methods which are of the most value in working empirically with texts and corpora, illustrating the various issues with numerous examples and moving from the most basic descriptive techniques (frequency counts and percentages) to decision-taking techniques (chi-square and z-score) and to more sophisticated statistical language models (Type-Token/Lemma-Token/Lemma-Type formulae, cluster analysis and discriminant function analysis).

Mathematical linguistics is basically a nonquantitative discipline. In a number of sciences the basic method of describing the properties of the objects being studied is the establishment of correlations between the quantities characterizing these objects. The situation seems to be different in linguistics. The most important characteristics of language do not have a quantitative nature, i.e. they are not quantities; the linguistic experiment usually does not entail measurements. It is not at all by chance that the main achievements of

linguistics, obtained over the entire period of its existence, have not been formulated as quantitative statements; in general, fundamental ‘nonquantitativeness is typical for overwhelming majority of linguistic works. Such a situation is connected with the very nature of the object of linguistics – with the properties of natural languages – and therefore appears quite normal.

The difference between qualitative and quantitative corpus analysis as the term themselves imply, is that in qualitative research no attempt is made to assign frequencies to the linguistic features which are identified in the data. Where as in quantitative research they classify features, count them, and even construct more complex statistical models in an attempt to explain what is observed, in quantitative research the data are used only as a basis for identifying and describing aspects of usage in the language and to provide real-life examples of particular phenomena. Schemied observes that, a stage of qualitative research is often a precursor for quantitative analysis, since, before linguistic phenomenon are classified and counted, the categories for classification must first be identified. Aim of qualitative analysis is a complete, detailed description. No attempt is made to assign frequencies to the linguistic features which are identified in the data, and rare phenomena receives (or should receive) the same amount of attention as more frequent phenomena. Qualitative analysis allows for fine distinctions to be drawn because it is not necessary to shoehorn the data into a finite number of classifications. Ambiguities, which are inherent in human language, can be recognised in the analysis. For example, the word "red" could be used in a corpus to signify the colour red, or as a political categorisation (e.g. socialism or communism). In a qualitative analysis both senses of *red* in the phrase "the red flag" could be recognised. The main disadvantage of qualitative approaches to corpus analysis is that their findings can not be extended to wider populations with the same degree of certainty that quantitative analyses can. This is because the findings of the research are not tested to discover whether they are statistically significant or due to chance.

In contrast to qualitative analysis, the quantitative analysis of a sampled corpus does allow for its findings to be generalised to a larger population, and furthermore, it means that direct comparisons can be made between two corpora, at least so long as valid sampling and significance techniques have been used. Thus, quantitative analysis allows us to discover which phenomena are likely to be genuine reflections of the behaviour of a language or variety, and which are merely chance occurrences. The more basic task of just looking at a single language variety allows one to get a precise picture of the frequency and rarity of particular phenomena, and thus their relative normality or abnormality. For statistical purposes, classifications have to be of the hard-and-fast (so-called "Aristotelian" type). An item either belongs to class *x* or it doesn't. So in the example about the phrase "the red flag" we would have to decide whether to classify "red" as "politics" or "colour". As can be seen, many linguistic terms and phenomena do not therefore belong to simple, single categories: rather they are more consistent with the recent notion of "fuzzy sets" as in the *red* example.

Quantitative analysis is therefore an idealisation of the data in some cases. Also, quantitative analysis tends to sideline rare occurrences. To ensure that certain statistical tests (such as chi-squared) provide reliable results, it is essential that minimum frequencies are obtained - meaning that categories may have to be collapsed into one another resulting in a loss of data richness.

Chomsky criticised corpus data as being only a small sample of a large and potentially infinite population, and that it would therefore be skewed and hence unrepresentative of the population as a whole. This is a valid criticism, and it applied not just to corpus linguistics but to any form of scientific investigation which is based on sampling. However, the picture is not as drastic as it first appears, as there are many safeguards which may be applied in sampling to ensure maximum representativeness. First, it must be noted that at the time of Chomsky's criticisms, corpus collection and analysis was a long and pains-taking task, carried out by hand, with the result that the finished corpus had to be of a manageable size for hand analysis. Although size is not a guarantee of representativeness, it does enter significantly into the factors which must be considered in the production of a maximally representative corpus. Thus, Chomsky's criticisms were at least partly true at the time of those early corpora. However, today we have powerful computers which can store and manipulate many millions of words. The issue of size is no longer the problem that it used to be. Random sampling techniques are standard to many areas of science and social science, and these same techniques are also used in corpus building. But there are additional caveats which the corpus builder must be aware of. Biber (1993) emphasises that we need to define as clearly as possible the limits of the population which we wish to study, before we can define sampling procedures for it. This means that we must rigourously define our sampling frame - the entire population of texts from which we take our samples.

This is the most straight-forward approach to working with quantitative data. Items are classified according to a particular scheme and an arithmetical count is made of the number of items (or tokens) within the text which belong to each classification (or *type*) in the scheme. For instance, we might set up a classification scheme to look at the frequency of the four major parts of speech: noun, verb, adjective and adverb. These four classes would constitute our types. Another example inolves the simple one-to-one mapping of form onto classification. In other words, we count the number of times each word appears in the corpus, resulting in a list which might look something like:

abandon:5
abandoned:3
abandons:2
ability:5
able:28
about:128
etc.....

More often, however, the use of a classification scheme implies a deliberate act of categorisation on the part of the investigator. Even in the case of word frequency analysis, variant forms of the same lexeme may be lemmatised before a frequency count is made. For instance, in the example above, *abandon*, *abandons* and *abandoned* might all be classed as the lexeme ABANDON. Very often the classification scheme used will correspond to the type of linguistic annotation which will have already been introduced into the corpus at some earlier stage. An example of this might be an analysis of the incidence of different parts of speech in a corpus which had already been part-of-speech tagged

The idea of collocations is an important one to many areas of linguistics. Khellmer (1991) has argued that our mental lexicon is made up not only of single words, but also of larger phraseological units, both fixed and more variable. Information about collocations is important for dictionary writing, natural language processing and language teaching. However, it is not easy to determine which co-occurences are *significant* collocations, especially if one is not a native speaker of a language or language variety. Given a text corpus it is possible to empirically determine which pairs of words have a substantial amount of "glue" between them. Two of the most commonly encountered formulae are: mutual information and the Z-score. Both tests provide similar data, comparing the probablities that two words occur together as a joint event (i.e. because they belong together) with the probability that they are simply the result of chance. For example, the words *riding* and *boots* may occur as a joint event by reason of their belonging to the same multiword unit (*riding boots*) while the words *formula* and *borrowed* may simply occur because of a one-off juxtaposition and have no special relationship. For each pair of words, a score is given - the higher the score the greater the degree of collocality. Mutual information and the Z-score are useful in the following ways:

- They enable us to extract multiword units from corpus data, which can be used in lexicography and particularly specialist technical translation.

- We can group similar collocates of words together to help to identify different senses of the word. For example, *bank* might collocate with words such as *river*, indicating the landscape sense of the word, and with words like *investment* indicating the financial use of the word.

- We can discriminate the differences in usage between words which are similar. For example, Church et al (1991) looked at collocations of *strong* and *powerful* in a corpus of press reports. Although these two words have similar meanings, their mutual information scores for associations with other words revealed interesting differences. *Strong* collocated with *northerly, showings, believer, currents, supporter* and *odor*, while *powerful* collocated with words such as *tool, minority, neighbour, symbol, figure, weapon* and *post*. Such information about the delicate differences in collocation between the two words has a potentially important role, for example in helping students who learn English as a foreign language.

Conclusion

John Lyons discusses about the importance of statistical structure in the language analysis. He points out that the linguistic and statistical principle stands in contradiction. The statistical considerations are essential to an understanding of the operations and development of languages. But at the same time it is impossible to calculate precisely the information carried by linguistic units in actual utterances. This apparent contradiction is resolved by recognizing that linguistic theory, at the present time at least, is not and cannot, be concerned with the production and understanding of utterances in their actual situations of use, but with the structure of sentences considered in abstractions from the situations in which actual utterances occur.

But, the quantitative linguists may claim that the limitations pointed out by Lyons (1968, 81-98) could be overcome by quantitative analysis based on corpus. Corpus linguistics is of course is not a branch of linguistics in the same sense as syntax, semantics, sociolinguistics, and so on. All of these disciplines concentrate on describing/explaining some aspect of language use. Corpus linguistics is in contrast is a methodology rather than an aspect of language requiring explanation or description. A corpus-based approach can be taken to many aspects of linguistic enquiry.

Reference:

Biber, D. 1993. Co-occurrence patterns among collocations: a tool for corpus based lexical language acquisition. Computational Linguistics 19(3):549-56.
Gladkij, A.V. & Mel'cuk, I.A. 1969. Elements of Mathematical Linguistics. Berlin: Mouton Publishers.
Herdan, G. 1964. Quantitative Linguistics. London: Butterworths.
Kuretzschmar, W.A. & Schneider, E.W. 1996. Introduction to Quantitative Analysis of Linguistic Survey Data. London: SAGE Publications.
Lyons, John. 1968. Introduction to Theoretical Linguistics. Cambridge: Cambridge University Press.
McEnery, T. & Wilson, A. 1996. Corpus Linguistics. Edinburgh: Edinburgh University Press.
Samarin, W.J. 1967. Field Linguistics: A Guide to Linguistic Field Work. New York: Holt, Rinehart and Wionston.